Going Beyond the Mean: Using Variances to Enhance Understanding of the Impact of Educational Interventions for Multilevel Models

Yadira Peralta *Mario Moreno* *Michael Harwell*
University of Minnesota

S. Selcen Guzey *Tamara J. Moore*
Purdue University

Variance heterogeneity is a common feature of educational data when treatment differences expressed through means are present, and often reflects a treatment by subject interaction with respect to an outcome variable. Identifying variables that account for this interaction can enhance understanding of whom a treatment does and does not benefit in ways that can inform and improve the treatment. Even in the absence of a treatment effect expressed through means studying variance heterogeneity offers insight into a treatment by identifying subject characteristics related to heterogeneity. This study illustrates four methods of modeling variance heterogeneity for data from a study of the impact of an engineering design-based STEM curriculum on student achievement with a focus on multilevel models

Introduction

Research in education at the K-12 (e.g., Fortus, Dershimer, Krajcik, Marx, & Mamlok-Naaman, 2004; Mehalik, Doppelt, & Schuun, 2008; Schnittka & Bell, 2011; Wendell & Rogers, 2013) and post-secondary level (e.g., Atadero, Rambo-Hernandez, & Balgopal, 2015; Carberry & McKenna, 2014; Hsiung, 2012; Lawton et al., 2012; Van Meter et al., 2016) often examines intervention (treatment) effects that are designed to promote learning and achievement. Experimental and quasi-experimental designs are common, and educational studies

increasingly use multilevel models to analyze data in which the means of treatment and control conditions are compared. If an intervention is effective then treatment and control condition means differ in ways that reflect the impact of the intervention.

When treatments are implemented in clustered settings such as students clustered within teachers and teachers clustered within schools group differences in dispersion is a common characteristic of the data (Raudenbush & Bryk, 1987). Snedecor and Cochran (1989), Raudenbush and Bryk (1987) and others have noted that when group means differ, group variances frequently differ in the same direction and that studying variances can provide important insights into the impact of an intervention. However, heterogeneity has usually been treated as a nuisance rather than a source of information about a treatment that should be studied (Bryk, 1977; Keppel, 1991; Raudenbush & Bryk, 1987), an unfortunate practice because variance heterogeneity is common in educational, behavioral, and psychological studies (Ruscio & Roche, 2012). In fact, there is often no reason to assume an intervention will be equally effective for all subjects in the treatment condition due to individual differences or other factors (Bryk & Raudenbush, 1988; Howell, 2013). In some cases, an explicit goal of an educational intervention is to reduce variability among students' outcomes (e.g., achievement gap reduction) based on the premise that successful schools should demonstrate high and relatively homogenous achievement (Kim & Choi, 2008). Therefore, studying variance heterogeneity should be central to data analysis.

The purpose of this study is to illustrate a series of methods for analyzing variance heterogeneity in multilevel models using data from an engineering design-based STEM curriculum program. Section 2 provides an argument of the need to analyze variances and what can be learned from

doing so, and Section 3 draws on the statistics literature to outline four methods for modeling variances. Section 4 describes a design-based engineering curriculum program and Section 5 applies the four methods to these data. Finally, the paper provides recommendations for using the four methods, how to interpret the results using data, and outlines implications of studying means and variances using the engineering curriculum study as an example.

The Need to Analyze Variances

Understanding the impact of an intervention reflected in means on an outcome variable can be enhanced by learning whether treatment and control variances are the same (homogeneity of variance) or different (heterogeneity of variance). Equal treatment and control condition variances imply that an intervention had a similar effect on students (i.e., the pattern of scores on an outcome was similar), and unequal variances that an intervention caused scores to bunch together (students responded similarly to an intervention) or spread out (students varied substantially in their response to the intervention). In both cases studying the pattern of variances can enhance understanding of who an intervention does and does not benefit in ways that can inform and improve the intervention.

General linear model-based analyses of means typically assume that samples come from populations sharing a common (error) variance; otherwise variances are heterogeneous. These analyses typically rely on traditional t-tests and F-tests that depend on data satisfying assumptions of independence, normality, and homogeneity of variance (Kutner, Neter, Nachtsheim, & Wasserman, 1996). There are numerous examples in the educational research literature of using t-tests to examine mean differences due to an intervention (e.g., Atadero et al., 2015; Fortus et al., 2004; Kollöffel & de Jong, 2013; Mehalik et al., 2008; Schnittka &

Bell, 2011; Wendell & Rogers, 2013) and/or F-tests (e.g., Van Meter et al., 2016). However, the homogeneity of variance assumption is rarely checked. For instance, from the educational research literature cited above only one article reported to have tested such assumption (Van Meter et al., 2016).

Analysis of Variance in Multilevel Models

A randomized cluster design in which students are clustered within classrooms is employed which leads to a multilevel model of the form:

$$Y_{ij} = \beta_{0j} + \sum_{q} \beta_{qj} X_{qij} + e_{ij} \quad 0 \qquad \text{(student model) (1)}$$

$$\beta_{pj} = \gamma_{\mathrm{p0}} + \sum_{r} \gamma_{\mathrm{pr}} W_{rj} + u_{pj} \qquad \text{(classroom model) (2)}$$

In equations (1) and (2) Y_{ij} is the outcome of the i-th student in the j-th classroom, β_{0j} is the intercept of the j-th classroom (j = 1,2,…,J), β_{qj} is the slope capturing the impact of the q-th student-level predictor X_{qi} which often represents a control variable, e_{ij} is a normally distributed student-level residual $e_{ij} \sim N(0, \sigma_j^2)$, β_{pj} is the p-th regression coefficient (p = 0, 1, 2, …, Q) for the j-th classroom, γ_{p0} is a classroom-level intercept, γ_{pr} is a slope capturing the impact of the classroom-level predictor W_{rj}, and u_{pj} is a normally distributed residual for the classroom model (Raudenbush & Bryk, 2002).

Several authors have argued variance heterogeneity in hierarchical (multilevel) models should be studied (Kim & Choi, 2008; Kim & Seltzer, 2011; Leckie, French, Charlton, & Browne, 2014). Consider a two-level randomized cluster design in which students (level 1) are nested within

classrooms (clusters, level 2) that is represented statistically in equations (1) and (2), and assume classrooms are assigned at random to a treatment or control condition. Assume Y is an outcome variable measuring students' achievement and both student (e.g., gender, race, i.e., X_{qi} in equation (1)) and classroom predictors (e.g., treatment indicator, percentage of English language learners, i.e., W_{rj} in equation (2)) appear in the model. Variance heterogeneity is commonly conceived as the result of non-modeled interaction effects of student characteristics with treatment (Bryk & Raudenbush, 1988; Kim & Seltzer, 2011), i.e., student characteristics × treatment interactions are present in the data and have not been taken into account. Such interactions can also occur between treatment and classroom characteristics and efforts to explicitly model interactions between treatment and covariates at both level 1 and 2 have been made (e.g., Mayer, Nagengast, Fletcher, & Steyer, 2014; Pituch, 2001; Plewis & Hurry, 1998). Non-modeled student- and classroom-level characteristics x treatment effects are not the only sources of variance heterogeneity, for example, measurement error could cause unequal variances (Bryk & Raudenbush, 1988). Nonetheless, ignoring unequal variances may lead to biased estimates of treatment effects or to incorrect or incomplete interpretations of mean (fixed) effects (Bryk & Raudenbush, 1988; Mayer et al., 2014).

A deeper understanding of the impact of a treatment on achievement is possible by examining both means and variances of Y. The latter reflect error variances estimated for each classroom ($\widehat{\sigma}_j^2$, j=1, 2,…, J classrooms) that represent variation in Y after student predictors have been taken into account. The premise is straightforward: We desire a treatment that is effective for all treatment students which implies these students benefit in a similar fashion from exposure to the treatment; in this case the treatment

condition mean would be larger, and the classroom residual variances smaller than those of the control condition (i.e., student achievement variability around classroom means is smaller for the treatment condition). An intervention that increases the treatment condition mean but produces larger variances compared to the control condition implies that, on average, the treatment is effective but treatment students do not benefit equally from exposure to the intervention.

In multilevel models the relationship between treatment and Y is assessed with a fixed effect, for example, a classroom-level slope (denoted by γ_{01} when treatment is considered the classroom predictor W_{1j} in equation (2)) capturing the impact of the treatment on Y. A statistical test of $\hat{\gamma}_{01}$ yields two possible results: (a) $\hat{\gamma}_{01} \neq 0$ meaning there is a treatment effect, (b) $\hat{\gamma}_{01} = 0$ meaning there is no treatment effect. If case (a) holds the treatment and control Y-means differ (conditional on the model) and if $\hat{\gamma}_{01}$ (estimated treatment effect) > 0 the implication is that the treatment on average raised student scores. A pattern in which treatment classrooms also showed larger $\hat{\sigma}_j^2$ than control classrooms implies that some treatment students benefited more than others relative to treatment classroom Y-means compared to control classrooms, i.e., there is a student × treatment interaction. This pattern suggests one or more variables are responsible for the student × treatment interaction, and including these variables as predictors in a regression model in which $\hat{\sigma}_j^2$ or some function of $\hat{\sigma}_j^2$ serves as the outcome can deepen our understanding of the treatment effect. Table 1 lists

Table 1: Possible outcomes of analyzing mean and variance differences in multilevel models

Effect of treatment	Variance Heterogeneity		Variance Homogeneity
	$\hat{\sigma}^2_{j_T} < \hat{\sigma}^2_{j_C}$	$\hat{\sigma}^2_{j_T} > \hat{\sigma}^2_{j_C}$	$\hat{\sigma}^2_{j_T} = \hat{\sigma}^2_{j_C}$
Treatment is effective ($\hat{\gamma}_{01} > 0$)	Treatment students benefitted uniformly relative to treatment classroom Y-means compared to control classrooms. In this case, treatment had a homogenizing effect.	Some treatment students benefited more than others relative to treatment classroom Y-means compared to control classrooms.	Treatment on average raised treatment classroom Y-means and variability around classroom Y-means was similar.
o treatment effect ($\hat{\gamma}_{01} = 0$)	Although the treatment/control mean difference was not significant, treatment tended to have a homogenizing effect on Y scores.	Although the treatment/control mean difference was not significant, treatment students did not respond uniformly. Some treatment students benefitted more than others as reflected in variation in Y scores about the classroom Y-means.	Failure of the treatment to raise scores was consistent across student and teacher characteristics.

Note. **$\hat{\sigma}^2_{j_T}$: Estimated residual variance associated with treatment classrooms. $\hat{\sigma}^2_{j_C}$: Estimated residual variance associated with control classrooms. *Y* denotes the outcome variable of interest.**

all possible outcomes of analyzing mean and variance differences in multilevel models.

Methods for Studying Variances

Several methods for analyzing variance heterogeneity in multilevel settings have been proposed in the statistics literature. Raudenbush and Bryk (1987) introduced the use of a two-level hierarchical model along with a log-transformation of residual variances to identify variables related to differences in residual variances across level 2 clusters for an outcome variable. These authors first estimate a standard two-level model (e.g., students within classrooms), and then apply a transformation to level-1 (within-classroom) residual variances involving the logarithmic function. Finally, assuming normality, they fit a single-level linear regression model to the log-transformed residual variances using classroom-level (level 2) predictors. Raudenbush and Bryk (1987) implemented this procedure in their Hierarchical Linear Modeling (HLM) software but only level-1 predictors can be used to model within-cluster variability in the current version [Version 7] (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). This limits the ability to identify classroom-related factors that might impact the variance of Y. For the interested reader, Leckie et al. (2014) provide an extensive review of a series of model extensions that have been proposed in the literature to analyze unequal within-cluster variances in multilevel models.

The presence of variance heterogeneity has often triggered the use of a variance-stabilizing transformation for *Y* followed by a test of mean differences on the transformed data (Bryk & Raudenbush, 1988; Howell, 2013). For instance, Kim and Seltzer (2011) proposed a single-level analysis of log-transformed residual variances obtained from

the estimation of a two-level hierarchical model. Kim and Choi (2008) proposed an alternative to the log-transformation in the dispersion model by modeling the square root of the within-cluster residual variance (SD) as a function of classroom-level predictors. Unfortunately successfully interpreting the results in the transformed scale may be challenging (Firth, 1988; Howell, 2013), and transforming the (previously transformed) data back to their original scale to enhance interpretation can be problematic because estimated differences among means can be reversed in the original scale (Grissom, 2000).

Alternatively, generalized linear models (GLMs) can be used to directly model the residual variance. The generalized modeling framework subsumes a variety of distributional assumptions for the outcome variable and provides maximum likelihood estimates of the parameters of a regression model (McCullagh & Nelder, 1989; Neuhaus & McCulloch, 2011). A gamma model is considered part of the family of GLMs, and the gamma distribution is particularly useful when modeling positively skewed data (McCullagh & Nelder, 1989), such as residual variance. Another advantage of GLMs is the possibility of a straightforward interpretation of the results in the original scale (Firth, 1988; McCullagh & Nelder, 1989). Thus the use of GLMs to directly model residual variances represents an important option to consider when analyzing heterogeneity.

The literature reviewed above directly provides three methods for modeling variance heterogeneity for multilevel data (Methods 1, 2, 3), and indirectly the foundation for a new method which we propose (Method 4). These four methods share the goal of studying variability to deepen understanding of a treatment effect but employ different statistical procedures. In all methods $\widehat{\sigma}_j^2$ are computed using ordinary least squares (OLS).

Method 1

The first method allows the impact of student and classroom predictors (including treatment) on variability to be modeled using a multilevel approach. Equations (1) and (2) define the multilevel model being estimated. The HLM7 software (Bryk, Raudenbush, & Congdon, 2011) allows unequal $\hat{\sigma}_j^2$ to be modeled using student-level predictors in conjunction with equations (1) and (2):

$$\mathrm{Ln}\left(\hat{\sigma}_{ij}^2\right) = \alpha_0 + \sum_j \alpha_j C_{ij} \qquad (3)$$

where ln represents the natural log, α_0 an intercept, and α_j a slope. If residuals (e_j in equation (1)) are normally-distributed, then $\ln \hat{\sigma}_j^2$ is approximately normally-distributed with variance $v_j = \frac{2}{df}$ (df = error degrees of freedom) which in a level 1 (student) regression model is $v_j = \frac{2}{n_j - Q - 1}$ (Raudenbush & Bryk, 1987), n_j = cluster sample size, Q = number of student predictors. C_{ij} in equation (3) represents student predictors used to account for variance heterogeneity.

It is important to emphasize that the same student predictors could be used in the level 1 model in equation (1) as well as in equation (3). For example, if gender is a significant predictor at level 1 with an estimated slope of 5 we would conclude that the Y-means of males and females differ by 5 units (conditional on the model). In this case $\hat{\sigma}_j^2$ have had the effects of gender removed in terms of the average effect of gender on Y. However, a slope of 5 tells us nothing about the ability of gender to predict variability of $\hat{\sigma}_j^2$. Including gender in equations (1) and (3) can provide information about whether means and variances differ across males and females. In theory, level 2 (classroom) predictors such as treatment could also be included to explain variance

heterogeneity. As noted earlier HLM7 limits equation (3) to level 1 predictors.

Method 2

Method 2 allows the impact of classroom predictors (including treatment) on variability to be modeled with a traditional normal-theory-based single-level regression. Initially a linear model (i.e., equation (1)) is fitted to each cluster (classroom) and $\ln \hat{\sigma}_j^2$ are computed. The $\ln \hat{\sigma}_j^2$ are then analyzed using a single level, weighted least squares regression with classroom predictors and weights v_j^{-1} defined in Method 1 —weights capture differences in classroom sample sizes. Because $\hat{\sigma}_j^2$ is computed independently for each classroom using OLS, Method 2 can be performed without any reference to multilevel modeling or multilevel software. For example, using R statistical software (R Core Team, 2013) we would fit the same level 1 regression model to each classroom using OLS via the dplyr package in R and obtain $\hat{\sigma}_j^2$ and then apply the natural logarithm to $\hat{\sigma}_j^2$

Method 3

Method 3 also allows the impact of classroom predictors including treatment on variability to be modeled with a single-level regression using a gamma model which is frequently recommended for continuous nonnegative data (Firth, 1988; McCullagh & Nelder, 1989). The within-classroom variances are again obtained by fitting a linear model for each classroom as in Methods 1 and 2. However, in this case the residual variance is modeled directly using a GLM (available in R statistical software via the glm function). Specifically, the GLM can be written as follows:

$$\hat{\sigma}_j^2 \sim \text{Gamma}(\mu_j, \phi)$$
$$g(\mu_j) = \eta_j \quad (4)$$
$$\eta_j = \beta_0 + \beta_1 W_j$$

where η_j represents a linear predictor, W_j is a classroom predictor, β_0 is an intercept, β_1 a slope, $g(\mu_j)$ represents the link function which in this case is the natural logarithm $g(\mu_j) = \ln(\mu_j)$, and ϕ represents the dispersion parameter. The link function connects the mean of the response variable (μ_j) with the linear predictor (η_j). Hence the outcome variable is not transformed to estimate the model, as in Method 2; rather the logarithm function is applied to the expected value of the outcome variable. Weights v_j^{-1} are also used in this method, where $v_j = \frac{2\sigma_j^2}{n_j - Q - 1}$ (Raudenbush & Bryk, 1987), n_j = cluster sample size, and Q = number of student predictors.

Method 4

The fourth method examines the impact of treatment on variability using a meta-analytic approach that can be used when classrooms, teachers, schools, etc. are matched. Matching is widely recommended as a way to control for pre-existing cluster differences and enhance causal arguments about a treatment (WWC, 2014). This produces a matched pair of treatment and control classrooms. The difference between $\ln \hat{\sigma}_j^2$ for each matched-pair is used to compute an effect size (δ) that serves as an outcome in a meta-analytic regression:

$$\delta_k = \ln \hat{\sigma}^2_{\text{treat(k)}} - \ln \hat{\sigma}^2_{\text{control(k)}} \quad (5)$$

In equation (5) $\ln \hat{\sigma}^2_{\text{treat (k)}}$ is the natural logarithm of the estimated within-classroom error variance of the treatment classroom within the k-th (k = 1,2,…, K) matched pair, and $\ln \hat{\sigma}^2_{\text{control(k)}}$ represents the same for the control classroom within the k-th matched pair. A test based on Raudenbush (1997) is then used to test variability among the effect sizes δ_k and the effect of moderators on δ_k

A Study of STEM Achievement

To illustrate the four methods for modeling variance heterogeneity we use data from a National Science Foundation (NSF), Mathematics and Science Partnership (MSP)-funded project. The project purpose is to increase student learning of engineering, science and mathematics concepts in Grades 4 - 8 using an engineering design-based approach to teacher professional development and curricular development. Treatment teachers teach curricular materials developed within the project that reflect State and National standards in STEM. In a three-week long summer workshop teachers developed the curricula and increased their understanding about a variety of science and mathematics concepts and learned about engineering and technology design. During the subsequent school year teachers then implemented STEM curricular units.

Teachers who agreed to participate in the study but did not participate in the professional development served as a "business as usual" control condition. Because of the hierarchical nature of the data, two-level (students within classrooms) models were used to examine the impact of the treatment. The main research question in the engineering design-based curriculum project asked was: In what ways does participation in the engineering design-based curriculum affect students' content knowledge in the STEM disciplines? A second important question was Does the treatment reduce

gaps in achievement among students by race, gender, and limited English proficiency (LEP) status? Both mean and variance differences can help answer these questions. Accordingly, we first present the traditional multilevel results focused on mean differences and then use project data to study variance heterogeneity using the four methods described earlier.

Population and Sample, Research Design, and Variables

The sampled population(s) of the STEM achievement study consisted of students and classrooms/teachers for grades 4-8 in a Midwest state. Treatment and control teachers were from three large school districts serving diverse student populations. Outcomes consisted of project-constructed assessments designed to capture achievement in engineering (Authors et al., 2015). Student achievement scores were reported in logits which are widely used in Rasch analyses of test data and estimate a student's proficiency on an outcome. Both treatment and control students took engineering assessments at the beginning and end of the engineering design-based unit in which these topics were covered. Thus both pretest and posttest data were available for these assessments, with posttest data serving as the outcome.

Student predictors consisted of gender (0 = male, 1 = female), race (Black, Asian, Hispanic, and White with the latter serving as a reference group), and the engineering pretest scores. Classroom predictors included treatment (treatment = 1, control = 0), years of teaching experience, years in current position, percentage of special education students, and percentage of LEP students.

The study also used matching to provide a sensitivity test of findings from the two-level multilevel model with control variables. Treatment and control teachers were initially matched using propensity scores (Dehejia & Wahba,

2002), then using the MatchIt R package (Sekhon, 2011) we perform "one-to-one" matching. The final sample used in this study consisted of about 2,300 students: 1,443 students corresponding to 17 treatment teachers, and 852 students to 17 control teachers (multiple sections of the same class taught by the same teacher were pooled into a single class).

Applying the Four Methods for Studying Variances to STEM Achievement Data

We first fitted the model specified in equations (1) and (2) with the above student and classroom predictors. The results in Table 2 show that engineering pretest is a significant predictor of engineering posttest. Notice that treatment was not a significant predictor of engineering posttest scores ($\hat{\gamma}_{01} = -.542$, p = .052), meaning that there was not a mean difference in engineering posttest scores between treatment and control conditions (conditional on the model). However, the Bartlett test of homogeneity of variances (available within HLM7) was used to detect heterogeneity and was statistically significant ($p < .05$). Thus, there is a significant difference in residual variances across the J classrooms. This signals that there is probably a student × treatment interaction, meaning that the treatment tended to produce scores that were bunched together or spread out (relative to control classrooms). Figure 1 presents box plots to illustrate the range of the residual variance for treatment and control classrooms, treatment classrooms tend to have slightly smaller residual variances that show a greater range than those for control classrooms.

Applying Method 1 to within-classroom residual variances (see Table 3) showed that (a) engineering pretest was not a significant predictor of $\ln\hat{\sigma}_j^2$ ($\hat{\alpha}_1 = .003$, p = .895)

suggesting that engineering pretest scores were unrelated to classroom variances, (b) females scored on average higher than males ($\hat{\gamma}_{20} = .150$, p = .002) but variability in posttest scores was the same across males and females ($\hat{\alpha}_2 = -.122$, p = .057), (c) Asian students scored on average lower than White students on the engineering posttest ($\hat{\gamma}_{30} = -.208$, p = .004) but their scores were less variable than White students ($\hat{\alpha}_3 = -.181$, p = .030). Thus, when investigating the effect of level 1 predictors on differences in variability we found the STEM engineering intervention produced a homogenizing effect for

Table 2: Estimation results of the multilevel model for the engineering posttest outcome

Fixed Effects	Coefficient	SE	*t*-ratio	*p*-value
	For Intercept Level 1, β_0			
Intercept level 2, γ_{00}	1.095	0.517	2.118	0.047*
Treatment, γ_{01}	-0.542	0.263	-2.064	0.052
Level, γ_{02}	-0.231	0.222	-1.040	0.311
LEP, γ_{03}	0.114	0.097	1.172	0.256
Special education, γ_{04}	-0.082	0.082	-0.999	0.330
Years teaching experience, γ_{05}	-0.089	0.109	-0.811	0.427
Years current position, γ_{06}	-0.026	0.118	-0.224	0.826
Years current school, γ_{07}	-0.163	0.132	-1.234	0.232
Gender of teacher, γ_{08}	-0.302	0.212	-1.424	0.170
Quality of curriculum unit, γ_{09}	-0.038	0.137	-0.280	0.782
Type of eng. integration, γ_{010}	0.217	0.162	1.338	0.196
RTOP, γ_{011}	0.292	0.162	1.810	0.085
	For Engineering Pre Score slope, β_1			
Intercept level 2, γ_{10}	0.588	0.134	4.402	0.000*

Treatment, γ_{11}	0.211	0.067	3.168	0.005*
Level, γ_{12}	0.182	0.069	2.632	0.016*
LEP, γ_{13}	-0.064	0.026	-2.501	0.021*
Special education, γ_{14}	-0.070	0.021	-3.373	0.003*
Years teaching experience, γ_{15}	0.058	0.035	1.636	0.117
Years current position, γ_{16}	-0.047	0.032	-1.470	0.157
Years current school, γ_{17}	0.015	0.038	0.406	0.688
Gender of teacher, γ_{18}	-0.059	0.054	-1.084	0.292
Quality of curriculum unit, γ_{19}	0.084	0.035	2.409	0.026*
Type of eng. integration, γ_{110}	0.017	0.044	0.388	0.702
RTOP, γ_{111}	-0.114	0.047	-2.421	0.025*
For Gender of Student slope, β_2				
Intercept level 2, γ_{20}	0.150	0.048	3.101	0.002*
For Asian slope, β_3				
Intercept level 2, γ_{30}	-0.208	0.072	-2.908	0.004*
For Hispanic slope, β_4				
Intercept level 2, γ_{40}	1.259	0.730	1.724	0.100

Table 2: Estimation results of the multilevel model for the engineering posttest outcome (cont.)

Fixed Effects	Coefficient	SE	*t*-ratio	*p*-value
Treatment, γ_{41}	0.041	0.318	0.129	0.899
Level, γ_{42}	-0.136	0.299	-0.455	0.653
LEP, γ_{43}	-0.083	0.125	-0.662	0.515
Special education, γ_{44}	-0.049	0.108	-0.455	0.653
Years teaching experience, γ_{45}	-0.152	0.149	-1.020	0.320
Years current position, γ_{46}	-0.012	0.157	-0.074	0.942
Years current school, γ_{47}	0.082	0.174	0.472	0.642
Gender of teacher, γ_{48}	-0.322	0.234	-1.377	0.184
Quality of curriculum unit, γ_{49}	0.064	0.165	0.385	0.704
Type of eng. integration, γ_{410}	-0.501	0.218	-2.302	0.032*
RTOP, γ_{411}	-0.009	0.210	-0.045	0.965
	For Black slope, β_5			
Intercept level 2, γ_{50}	0.272	0.570	0.477	0.638
Treatment, γ_{51}	-0.082	0.278	-0.295	0.771
Level, γ_{52}	-0.476	0.272	-1.750	0.095

LEP, γ_{53}	-0.049	0.104	-0.467	0.645
Special education, γ_{54}	-0.070	0.087	-0.808	0.429
Years teaching experience, γ_{55}	-0.178	0.137	-1.306	0.207
Years current position, γ_{56}	0.067	0.142	0.470	0.643
Years current school, γ_{57}	0.013	0.160	0.082	0.936
Gender of teacher, γ_{58}	-0.035	0.198	-0.176	0.863
Quality of curriculum unit, γ_{59}	0.127	0.148	0.859	0.401
Type of eng. integration, γ_{510}	-0.063	0.180	-0.351	0.729
RTOP, γ_{511}	0.117	0.186	0.629	0.536
Random Effect	SD	Variance Component	Chi-square	*p*-value
Intercept level 1, U_0	0.381	0.145	177.302	0.000
Engineering Pre Score slope, U_1	0.023	0.001	36.819	0.009
Hispanic slope, U_4	0.084	0.007	18.735	>.500
Black slope, U_5	0.187	0.035	29.832	0.054

Note. *** = statistically significant (p < .05). Gender is coded 1 = female, 0 = male; Treatment is coded 1= treatment, 0 = control; Level is coded 1 = middle school, 0 = elementary; The Reformed Teaching Observation Protocol (RTOP) is coded into three quality of teaching categories 0 = Low, 1 = Medium, 2 = High; Quality of curriculum unit is coded 0 = Not Present, 1 = Weak, 2 = Adequate, 3 = Good, 4 = Excellent; Type of engineering integration is coded 0 = add-on, 1 = implicit, 2 = explicit. Deviance = 6020.691; number of estimated parameters = 66.**

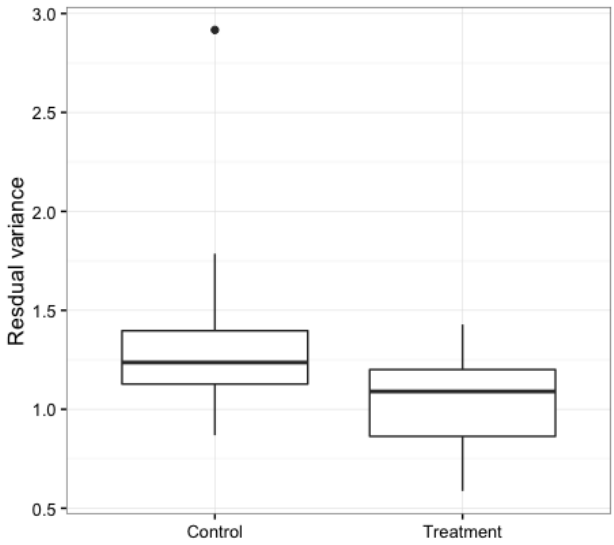

Figure 1. Boxplot of OLS residual variance by treatment condition for all classrooms.

Asian students compared with White students. That is, the logarithms of the residual variance estimates were smaller for Asian students than for their White counterparts indicating the responses of Asian students to the intervention were more similar than those of White students; the latter suggest more scores varied more (higher or lower than the mean) for White students compared to the former. This finding could guide efforts to better understand why there was greater variation in mastery of the material among White students. Changes in curriculum program or in teachers' preparation could help assure patterns of scores do not differ by race, an important instructional goal

Table 3: Estimation results of multilevel modeling of log-transformed residual variances

Parameter	Estimate	SE	Z	*p*-value
Intercept, α_0	0.198	0.058	3.403	0.001
Pretest, α_1	0.003	0.023	0.133	0.895
Gender, α_2	-0.122	0.064	-1.900	0.057
Asian , α_3	-0.181	0.083	-2.167	0.030
Hispanic, α_4	0.095	0.110	0.861	0.389
Black, α_5	0.038	0.088	0.437	0.662

Method 1 offers insight into the effect of level 1 predictors on variance heterogeneity whereas Methods 2 and 3 allow the impact of classroom-related factors on variability in the outcome variable to be investigated, and these are illustrated next.

For Method 2 we fitted a single regression model with one predictor to explain variance heterogeneity between treatment and control classrooms. This model adopted the following mathematical representation:

$$\ln\hat{\sigma}_j^2 = \beta_0 + \beta_1 W_j + e_j, \qquad (6)$$

where $\ln\hat{\sigma}_j^2$ represents the outcome variable, W_j is the predictor representing the treatment condition for each classroom (1 = treatment, 0 = control), β_0 is the intercept of the model and β_1 is the slope for the treatment predictor which captures the impact of engineering curriculum implementation on residual variances. We are interested in β_1 because this parameter indicates whether the variances of treatment classrooms were larger, equal to, or smaller than those of control classrooms. Table 4 shows that the natural logarithm of the variance in treatment classrooms was on average significantly smaller than the natural logarithm of the

variance in control classrooms ($\hat{\beta}_1 = -.226$, p = .022). This suggests that the engineering curriculum implementation tended to produce a homogenizing effect across students in treatment classrooms compared to students in control classrooms. In other words, treatment did not on average raise posttest scores (since there was not a treatment effect resulting from equations (1) and (2)) but treatment seemed to produce less within-classroom variability compared to control classrooms. This result is particularly important when programs are designed to decrease variability among students such as those aiming to reduce achievement gaps. In this context, differences in variability would indicate that even though on average treatment students did not present higher scores, students reacted to the intervention in similar ways.

Method 3 used the generalized modeling framework to investigate the relationship between residual variances and the treatment condition. The residual variance is a non-negative positively skewed variable (skew = 2.169, kurtosis = 7.188) that can be directly modeled by assuming a gamma distribution. The model is represented in equation (4) in which $\hat{\sigma}_j^2$ is the outcome variable and W_j is a dichotomous variable representing the treatment condition for each classroom. Results in Table 4 show that the variance in classrooms where the treatment was implemented was significantly smaller than in classrooms under the business as usual condition ($\hat{\beta}_1 = -.269$, p = .007), indicating that treatment had a homogenizing effect. The results for Methods 2 and 3 have the same practical interpretation but the advantage of Method 3 is the possibility of discussing differences in dispersion in the variance scale (as shown below) instead of in the log-variance scale (as in Method 2).

Table 4: Estimation results of Methods 2 and 3

Variable	Estimate	SE	*t*-ratio	*p*-value
Method 2: single level regression of log-transformed $\hat{\sigma}_j^2$				
(Intercept)	0.220	0.075	2.934	0.006
Treatment	-0.226	0.094	-2.414	0.022
Method 3: generalized gamma linear model of $\hat{\sigma}_j^2$				
(Intercept)	0.155	0.086		0.071
Treatment	-0.269	0.099		0.007

Note. **Method 2: R-square = 0.158. Method 3: Deviance = 66.667, AIC = -59.682.**

As previously mentioned one of the advantages of the GLM framework is the possibility of a straightforward interpretation of the results in the original scale. Since the link function in this model is the natural logarithm $g(\mu_j) = \ln(\mu_j) = \eta_j = \beta_0 + \beta_1 W_j$, then $\mu_j = \exp(\beta_0 + \beta_1 W_j)$. The estimated residual variance for the control condition ($W_j = 0$) equals $\exp(\hat{\beta}_0) = \exp(.155) = 1.167$ ($\hat{\beta}_0 = .155$, p = .071), and for the treatment condition ($W_j = 1$) equals $\exp(\hat{\beta}_0 + \hat{\beta}_1) = \exp(.155 - .269) = .892$. Hence, the difference in variability between the treatment and control conditions is given by $\exp(\hat{\beta}_0 + \hat{\beta}_1) - \exp(\hat{\beta}_0) = -.275$, which means that the residual variance of the engineering scores in the treatment condition is .275 units smaller than in the control condition.

In the meta-analytic approach (Method 4) we estimated a unique effect size for each pair of matched classrooms by computing the difference in $\ln \hat{\sigma}_j^2$ of the treatment and control classrooms (see equation 5). In total we obtained 17-effect sizes and then estimated the average weighted effect size across pairs as -.215, which suggests that on average treatment classrooms were less variable than control classrooms. Our next step was to perform a

statistical test to examine variability among the effect sizes (Raudenbush, 1997), which was statistically significant (148.456, $p < .001$) and indicates that effect sizes were heterogeneous, meaning some classrooms showed similar variation in treatment-control variances whereas others showed greater variation. This finding suggest non modeled teacher or classroom variables (e.g., years of experience, class size) may be responsible. Notice the practical interpretation of this method is based on effect sizes rather than the residual variance as in Method 3 or in the log-residual variance as in Method 2.

In summary, a comprehensive investigation of treatment effects is possible by coupling mean-oriented fixed effects results of multilevel models with methods to analyze variance heterogeneity. The four methods illustrated in this study serve different purposes which should guide adoption of one or more of these methods. Method 1 focuses on examining the relationship between level 1 predictors and log-transformed residual variances, whereas Methods 2, 3 and 4 investigate the effect of the level 2 treatment condition (a classroom-level predictor) on variability of the outcome variable. Applying Method 1 to STEM achievement data showed Asian students presented less within-classroom variability (were more homogeneous given the model) than White students. Methods 2, 3 and 4 provided the same general conclusion for the STEM achievement data in that students in the treatment condition showed less variability in posttest scores than students in the control condition (given the model). Differences between the last three methods are due to the nature of the outcome variable: Method 2 employs a log-variance scale, Method 3 a within-classroom residual variance scale, and Method 4 uses effect sizes. The results of all methods provide insight into whether students reacted similarly or not, and if the latter should prompt additional investigation to identify sources of differences in variances.

Discussion

The current study described and illustrated four methods for investigating variance heterogeneity that can deepen understanding of a treatment using data from an engineering design-based STEM curriculum study. Using a traditional multilevel modeling approach we found that integrated STEM instruction was not a significant predictor of engineering posttest scores (i.e., treatment and control conditions had the same mean achievement given the model). However, there was variance heterogeneity across treatment and control classrooms. Method 1 analyzed the relationship between log-transformed residual variances and student level predictors, allowing the role of student characteristics on variability to be explored. For example, the STEM engineering integration intervention produced a difference in engineering posttest means among male and female students but no relationship between classroom variances and gender. The latter finding implies that male and female students responded similarly on the outcome and suggests that the curriculum program does not need to be modified to help ensure gender is unrelated to dispersion. On the other hand, the results also indicated that White students on average scored higher than Asian students on the engineering posttest and that Asian students showed less variability in scores (relative to classroom means) and White students more variability (relative to classroom means). Ideally White and Asian students would respond similarly to the curriculum program and these findings may suggest a need to examine the curriculum for clues about why differences in variability emerged.

Methods 2 and 3 use level 2 (classroom) predictors to explain variability in the outcome variable using a single-level regression model. Method 4 provides an additional option to

investigate residual variances that, to our knowledge, has not been used in the educational literature. Educational studies that allow the meta-analytic framework to be applied produce information about differences in effect sizes and their relationship with different predictors including treatment. Methods 2, 3 and 4 showed that students in the treatment condition generally demonstrated less dispersion in engineering posttest scores than students in the control condition.

Although Methods 2, 3 and 4 represent an option to investigate the effect of classroom level predictors and no special multilevel software is needed to implement them, there are important differences between the three methods worth noticing. The difference between Methods 2 and 3 lies in the modeling framework that accommodates the outcome variable under study. Method 2 is located under normal statistical theory and involves a nonlinear transformation (logarithmic) of the residual variance estimates. Method 3 deviates from the normality assumption and models the residual variance estimates directly by assuming a gamma distribution. The practical implication of these two choices is the scale of the results: It is possible to have an interpretation in the original scale when using Method 3 but interpretations of the results for Method 2 need to be in the logarithmic scale. Method 4 focuses on analyzing differences in effect sizes, that is, it examines whether classroom level predictors are related to differences in variability between matched pairs of treatment and control classrooms. Notice that matched pairs of classrooms, teachers, schools, etc. are necessary to implement Method 4, which may impose a significant data restriction.

In general, the four methods represent different options for the analysis of variance heterogeneity in educational studies that provide evidence of whether treatment and control conditions and/or student and teacher

characteristics account for heterogeneity. In conjunction with analyses of mean differences in multilevel models the results of modeling variance heterogeneity help to answer the research questions posed earlier in ways that enrich inferences about the impact of an educational intervention on students. We recommend Method 1 when the focus of research is on level 1 predictors only. When classroom predictors are of interest, researchers should turn to Methods 2, 3 or 4 with the choice depending on the researcher' preference for the outcome variable and interpretation of results.

References

Atadero, R. A., Rambo-Hernandez, K. E., & Balgopal, M. M. (2015). Using social cognitive career theory to assess student outcomes of group design projects in statics. *Journal of Engineering Education, 104*(1), 55–73. http://doi.org/10.1002/jee.20063

Bryk, A. S. (1977). Evaluating program impact: A time to cast away stones, a time to gather stones together. *New Directions for Program Evaluation, 1*, 31–58.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396–404. http://doi.org/10.1037/0033-2909.104.3.396

Carberry, A. R., & McKenna, A. F. (2014). Exploring student conceptions of modeling and modeling uses in engineering design. *Journal of Engineering Education, 103*(1), 77–91. http://doi.org/10.1002/jee.20033

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Method, 13*(2).

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics* (4th ed.). Pearson.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1), 151–161.

Firth, D. (1988). Multiplicative errors: Log-normal or gamma? *Journal of the Royal Statistical Society. Series B, 50*(2), 266–268.

Fortus, D., Dershimer, R. C., Krajcik, J., Marx, R. W., & Mamlok-Naaman, R. (2004). Design-based science and student learning. *Journal of Research in Science Teaching, 41(*10), 1081–1110. http://doi.org/10.1002/tea.20040

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*(1), 155–165. http://doi.org/10.1037//00

Gu, S. X., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405–420.

Harwell, M. (1997). An investigation of the Rodenbush (1988) test for studying variance heterogeneity. *The Journal of Experimental Education, 65*(2).

Harwell, M., Moreno, M., Phillips, A., Guzy, S.S., Moore, T.J. & Rhoehrig, G. H.. (2015). A study of STEM assessments in Engineering, Science and Mathmatics for elementary and middle school students. *School Science and Mathmatics, 115*(2), 66-74. Doi:10.111/ssm.12105.

Hettmansperger, P. T., & McKean, J. W. (1998). *Robust nonparametric statistical methods*. London, U.K.: Arnold.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed). Belmont, CA, USA: Wadsworth, Cengage Learning.

Hsiung, C.-M. (2012). The effectiveness of cooperative learning. *Journal of Engineering Education*, 101(1), 119–137. http://doi.org/10.1002/j.2168-9830.2012.tb00044.x

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ, USA: Prentice Hall.

Kim, J., & Choi, K. (2008). Closing the gap: Modeling within-school variance heterogeneity in school effect studies. *Asia Pacific Education Review, 9*(2), 206–220. http://doi.org/10.1007/BF03026500

Kim, J., & Seltzer, M. (2011). Examining heterogeneity in residual variance to detect differential response to treatments. *Psychological Methods, 16*(2), 192–208. http://doi.org/10.1037/a0022656

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (1996). *Applied linear statistical models* (4th ed.). New York, NY: McGraw-Hill.

Lawton, D., Vye, N., Bransford, J., Sanders, E., Richey, M., French, D., & Stephens, R. (2012). Online learning based on essential concepts and formative assessment. *Journal of Engineering Education, 101*(2), 244–287. http://doi.org/10.1002/j.2168-9830.2012.tb00050.x

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance-covariance components in two-level models. *Journal of Educational and Behavioral Statistics, 39*(5), 307–332. http://doi.org/10.3102/1076998614546494

Mayer, A., Nagengast, B., Fletcher, J., & Steyer, R. (2014). Analyzing average and conditional effects with multigroup multilevel structural equation models. *Frontiers in Psychology, 5*(APR), 1–16. http://doi.org/10.3389/fpsyg.2014.00304

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2nd ed.). London: Chapman and Hall.

Mehalik, M. M., Doppelt, Y., & Schuun, C. D. (2008). Middle-school science through design-based learning versus scripted inquiry: Better overall science concept learning. *Journal of Engineering Education, 97*(1), 71–85. http://doi.org/10.1002/j.2168-9830.2008.tb00955.x

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156 –166.

Neuhaus, J., & McCulloch, C. (2011). Generalized linear models. Wiley Interdisciplinary Reviews. *Computational Statistics, 3*(5), 407–413.

Peng, C., Harwell, M., Liou, S., & Ehman, L. (2006). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 31–78). Charlotte, NC: New Information Age Publishing.

Pituch, K. A. (2001). Using multilevel modeling in large-scale planned variation educational experiments: Improving understanding of intervention effects. *The Journal of Experimental Education, 69*(4), 347–372.

Plewis, I., & Hurry, J. (1998). A multilevel perspective on the design and analysis of intervention studies. *Educational Research and Evaluation, 4*(1), 13–26. http://doi.org/10.1076/edre.4.1.13.13014

Raudenbush, S. W. (1988). Estimating change in dispersion. *Journal of Educational Statistics, 13*(2), 148–171. http://doi.org/10.3102/10769986013002148

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics, 12*(3), 241–269. http://doi.org/10.1177/1362361308098514

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, S. H. C. (2011). HLM 7. Lincolnwood, IL: Scientific Software International, Inc.Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology, 8*(1), 1–11. http://doi.org/10.1027/1614-2241/a000034

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. http://doi.org/10.1037//1082-989X.7.2.147

Schnittka, C., & Bell, R. (2011). Engineering design and conceptual change in science: Addressing thermal energy and heat transfer in eighth grade. *International Journal of Science Education, 33*(13), 1861–1887. http://doi.org/10.1080/09500693.2010.529177

Sekhon, J. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software, 42(*7), 52. http://doi.org/10.1.1.335.7044

Serlin, R. C., & Harwell, M. R. (2004). More powerful tests of predictor subsets in regression analysis under nonnormality. *Psychological Methods, 9*(4), 492–509. http://doi.org/10.1037/1082-989X.9.4.492

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, Iowa, USA: Iowa State University Press.

Timm, N. H. (2002). *Applied multivariate analysis*. New York, NY: Springer.

Van Meter, P. N., Firetto, C. M., Turns, S. R., Litzinger, T. A., Cameron, C. E., & Shaw, C. W. (2016). Improving students' conceptual reasoning by prompting cognitive operations. *Journal of Engineering Education, 105*(2), 245–277. http://doi.org/10.1002/jee.20120

Wendell, K. B., & Rogers, C. (2013). Engineering design-based science, science content performance, and science attitudes in elementary school. *Journal of Engineering Education, 102*(4), 513–540. http://doi.org/10.1002/jee.20026

What Works Clearinghouse (2014*). Procedures and standards handbook*. Retrieved from http://ies.gov/ncee/wwc/